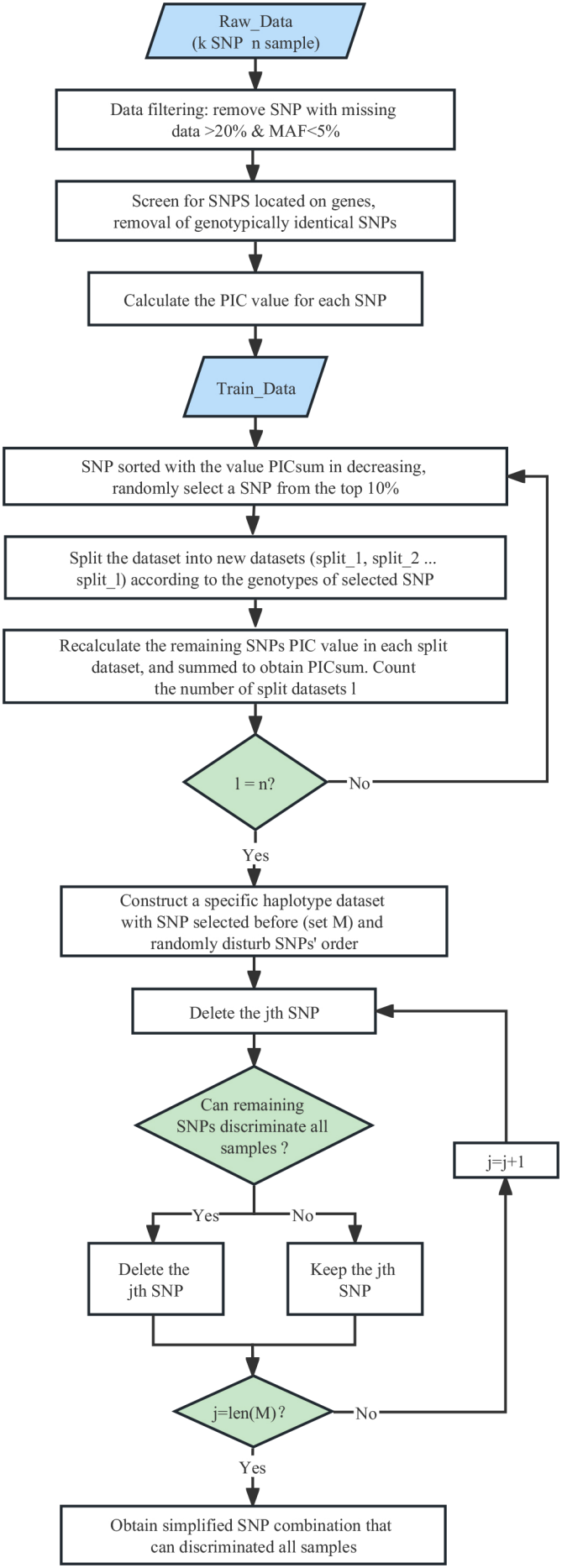

Raw_Data
(k SNP n sample)
Data filtering: remove SNP with missing data >20% & MAF<5%
Screen for SNPS located on genes, removal of genotypically identical SNPs
Calculate the PIC value for each SNP
Train_Data
SNP sorted with the value PICsum in decreasing, randomly select a SNP from the top 10%
Split the dataset into new datasets (split_1, split_2 ... split_l) according to the genotypes of selected SNP
Recalculate the remaining SNPs PIC value in each split dataset, and summed to obtain PICsum. Count the number of split datasets l
l = n?
No
Yes
Construct a specific haplotype dataset with SNP selected before (set M) and randomly disturb SNPs' order
Delete the jth SNP
Can remaining SNPs discriminate all samples ?
Yes
No
Delete the jth SNP
Keep the jth SNP
j=j+1
j=len(M)?
No
Yes
Obtain simplified SNP combination that can discriminated all samples